Modelling user’s preferences of public bus transportation using Neural Networks and GIS

R. Antunes¹, Y. Yamashita¹, A. Dantas², M. V. Lamar³

¹Master Program in Transportation, University of Brasilia, Brazil.
²Dept. of Civil Engineering, Nagoya Institute of Technology, Japan.
³Dept. of Electrical Engineering, Federal University of Paraná, Brazil.

Abstract

In Brazil, Public Bus Transportation (PBT) responds for the most part of displacements. However, due to low level of services provided by bus operators and the increase of informal/non-regulated transportation, there has been a sharp decrease on demand. In this context, planning agencies need now to focus on efforts towards comprehension and development of strategies to satisfy various necessities of PBT’s users. As a preliminary step in this direction, this work introduces the application of Neural Networks (NN) associated to a Geographical Information System (GIS) in order to process the modelling of user’s preferences in PBT. A case study was conducted in Taguatinga City, Federal District, Brazil. Results show the potential of the proposed model.

1 Introduction

In developing countries such as Brazil, PBT has been fundamental but it still lacks of a special treatment. Traditionally, PBT has been planned considering users as a homogeneous group without any regards to their particular characteristics and needs. Consequently, it has been noticed the decrease of PBT’s demand due to growing competition with informal modes such as vans, minibuses and motorcycles, which are more flexible and suitable to fulfill the complex needs of PBT’s users.

In order to revert this situation, planning agencies need to focus on efforts towards comprehension and development of strategies to accomplish needs of PBT’s users. Knowledge and comprehension on users’ characteristics and necessities and their relation with the dynamic of urban areas are fundamental.
Nevertheless, it is well known that this is not a simple task to be conducted since the nature of the variables involved is highly heterogeneous leading to many difficulties on their representation.

As a preliminary step in this direction and to model the preferences of PBT's users, we propose here the application of Neural Networks (NN) associated to a Geographical Information System (GIS). This model provides the forecasting of user's preferences expressed in terms of price, safety, frequency, comfort and speed based upon socio-economic and spatial data, which are acquired from GIS operations and then processed using NN. GIS-NN integration intends to take advantage of new computational techniques based upon georeferenced data and a non-linear classifier function, which provides a flexible and self-adaptive modelling. These characteristics can be decisive when solving geographical-related problems, since it has been shown [1] that due to their complexity it would be quite impossible to solve them if statistical regression analysis was used.

This paper is organized into five sections. After this introduction, we describe the main characteristics of PBT system in Brazilian cities. In the sequence, the proposed model is described on section three. In the fourth section, a case study in Taguatinga City, Federal District, Brazil is reported. Finally, we discuss the obtained results and general perspectives for future improvements of this research.

2 Public Bus Transportation: Brazilian context

PBT has an essential role in the urban structure of Brazilian cities. Along the years, it has been consolidated as the main transportation mode due to its flexibility and low costs on implementation [2]. According to NTU [3], 60% of daily trips in Brazilian cities, with more than 100,000 inhabitants, are performed by bus. Additionally, it is noticed that the most part of users has PBT as the only option on their displacements.

However, this scenario is rapidly changing as a consequence of low quality level provided by PBT system. As users are not satisfied with the system, that has ignored needs and basic requirements to maintain an acceptable level of transportation, sharp decreases on PBT's demand in the most important Brazilian metropolitan areas, increase on traffic congestion and pollution are now detected. On the other hand, informal transportation, which is a non-regulated service supplied by anyone with a motorized vehicle, is increasing its participation on modal split. This kind of service is totally irregular in the sense that drivers do not have any training, there is no safety and insurance and taxes are not obtained from its activities. Moreover, informal transportation is not concerned with frequency and traffic regulations, but only devoted to collect passengers.

Towards the development of strategies focusing on user's needs, they must be faced as "clients". Users have to be satisfied according to their characteristics, wishes and profiles. In this context, Marketing theories are instruments to be explored in the sense that PBT's market is analyzed under new paradigms that concentrate on clients' demands. In this direction, the transformation has to start
by identifying potential users and clarifying what they want from the system [4]. Based upon this initial knowledge, it will be possible to formulate and keep services for this new reality of PBT.

Few efforts in this direction have been observed to analyze user's preferences under different perspectives and verify their influence in PBT planning. Just recently, Martins [5] proposed a methodological framework towards the segmentation of PBT market, but it lacks of a general mathematical formulation to model the preferences and it did not incorporate spatial reality as a factor affecting PBT’s users and their decision making process.

In addition to this scenario of incipient theoretical background, obtainment of information on user’s preferences is decisive to correctly express PBT’s market. Due to narrow relations between users and urban environment (land use, traffic system, etc), analytical evaluations of these preferences must be performed considering spatial dependencies, as pointed out by Fisher [6] and Longley et al. [7]. Therefore, spatial information associated to socio-economic and behavioural data is fundamental to express the complex nature of urban problems such as the modelling of user’s preferences. However, costs involved on data collection are mostly prohibitive for planning agencies of developing countries, so models have to be conceived facing such a limitation.

3 User’s preferences modelling

In this section we describe a neural-spatial model to forecast attributes related to daily trips as part of the effort to obtain knowledge on user’s preferences. This model concentrates on the representation and incorporation of user’s characteristics, urban environment and its properties and relationship between the urban space and spatial location. These are main factors that have to be carefully incorporated in order to forecast user’s preferences of PBT due to their influence on user’s behaviour. Traditionally, user’s characteristics such as socio-economic data are obtained and treated through statistical regression, while urban related aspects are hardly analyzed and considered into transportation studies [8]. In this model, we intend to combine both and reach a comprehensive representation of PBT user’s preferences.

In this sense, we integrate theoretical fundamentals of Marketing, NN and GIS. From Marketing, we assimilated the necessity to treat users as clients and in the selection of socio-economic data to represent the preferences. Additionally, GIS is used to obtain, manipulate, analyze and create spatial information into the modelling process. Finally, NN is responsible for the achievement of a non-linear modelling function [9]. Both GIS and NN have been extensively and successfully applied to solve transportation problems as described by Stokes and Marucci [10] and Himanen et al. [11]. However, these applications have separately explored without taking fully advantage of their interrelated capabilities, but as suggested by Fisher [12] they have to work together.

In this sense, the modelling of preferences starts by defining the general assumptions to express user’s behaviour. So, a person (or a PBT’s user) commuting from his/her residential location, is described by two main groups of
characteristics: individual; and spatial. Through the computation of these characteristics, we assume that it is possible to forecast user’s preference related to attributes such as price, safety, frequency, comfort and speed. Individual characteristics represent the primary level of information. Generally, they can be divided into four groups: socio-economic, demographic, user’s habits and preferences related to trip’s attributes. On the other hand, spatial characteristics describe the urban environment (traffic system, land use) and the relationship between potential users, as well as the influence of location on their preferences (location and level of accessibility). They can be divided into two classes: urban environment; and spatial-locational. The former provides the representation of the region surrounding the potential user, while spatial-locational class provides a measurement of the position on the urban space.

Based on these general assumptions, we concentrate on the mathematical formulation of the NN, which is defined by Haykin [13] as a massively parallel distributed processor made up of simple processing units dedicated to the storage and use of experimental knowledge. In this modelling, we make use of the most applied NN structure that is a feedforward Multilayer Perceptron (MLP). In this NN structure, we have to establish the composition of the input and the output vectors needed to training and testing activities of this modelling [14]. Initially, we define the vector \( \vec{X} \) as

\[
\vec{X}_m = (SE_m, DM_m, HV_m, SV_m, US_m, LZ_m, NA_m)
\]

where:
- \( \vec{X}_m \) is the input vector for a sample \( m \);
- \( SE_m \), \( DM_m \), \( HV_m \), \( SV_m \), \( US_m \), \( LZ_m \), \( NA_m \) are, respectively, socio-economic, demographic, user’s habit, traffic system, land use, location and level of accessibility vectors for a sample \( m \).

On the other hand, the output vector \( \vec{Y} \) is a mathematical codification of a vector \( \vec{AV} \), as presented by equation 2. This vector contains priorities \( av^m \) of user’s preference for \( p \) attributes.

\[
\vec{AV}_m = (av^m_1, av^m_2, ..., av^m_p)
\]

Once \( \vec{AV} \) is obtained, equation 3 is applied as described next:

\[
y^m_k = \begin{cases} 
1, & \text{if } av^m_k = av^m_{\text{max}}, \text{ where } k \in \{1, 2, ..., p\} \\
0, & \text{otherwise}
\end{cases}
\]

where:
- \( y^m_k \) is the coded value for attribute \( av^m_k \);
- \( av^m_k \) is the attribute value \( k \) for sample \( m \); and
- \( av^m_{\text{max}} \) is the attribute considered most important for sample \( m \).
Finally, vector $\tilde{Y}$ is reached by applying the following equation:

$$\tilde{Y}_m = (y_1^m, y_2^m, ..., y_p^m)$$

(4)

Based on the definition of equations 1 and 4, figure 1 presents a general NN structure.

4 Case study

A case study was conducted in Taguatinga City, Federal District, Brazil, which comprehends 230 thousand people and occupies a 121,34 square km area. Located 25 Km from Brasilia, Taguatinga was planned to be a satellite city, but it has developed a large variety of independent activities from the main core city, which includes industrial and commercial areas and decisively contributes to the economy of this region.

The description of this case study is conducted in three phases: GIS database construction; NN simulations; and analysis of the results.

4.1 GIS database

Initially, it was conducted the diagnosis of PBT in Taguatinga city. In this sense, we obtained data previously collected by Martins [5]. This data set contains 276 samples related to demographic, socio-economic, user’s habits and preferences related to trip’s attributes (price, safety, frequency, comfort and speed). Next, in order to characterize spatial characteristics, we obtained a digital map database and a set of aerial photographs (scale 1:2,000, black&white) provided by the local development governmental agency (CODEPLAN), which was incorporated into a GIS database as shown in Figure 2.

Based upon this GIS database, firstly we digitized the traffic system (arterial, avenues, circulation axis, secondary streets and local streets). Then, following the United States Geological Service (USGS) classification system [15] and
Taco’s methodology [16], Land Use patterns (commercial, education, heath, services, leisure, sportive clubs and religion) were identified.  

![Digital database in the GIS software](image)

Figure 2: Digital database in the GIS software

Next, we processed spatial queries on the GIS database to obtain the data needed to process NN simulations. Then, 276 vectors containing individual and spatial characteristics were composed.

4.2 NN simulations

We defined four types of simulations to be conducted in order to evaluate different composition and treatment on the input vector. Specifically, our intention was to examine NN’s efficiency using spatial data (traffic system, land use, location and level of accessibility) as well as analyze codification and normalization pre-processing procedures. Table 1 shows the considered characteristics in each type of simulation.
The codification procedure is related to socio-economic, demographic and trip’s habits variables. In this procedure, individual characteristics are such as gender, age and income are transformed into a binary vector \( x^a_{m,z} \) for variable \( a \). This vector is composed by \( x^a_{m,z} \) values obtained through equation 5 defined as:

\[
x^a_{m,z} = \begin{cases} 
1, & \text{if } x^a_m = z \text{ where } a \in (1, 2, ..., w) \\
0, & \text{otherwise}
\end{cases}
\]  

(5)

where:
- \( x^a_m \) is the original value obtained from survey for variable \( a \) and sample \( m \);
- \( x^a_{m,z} \) is the coded value related to field \( z \);
- \( z \) is the total number of variables to be coded;

The normalization procedure is conducted to fit original values such as areas, extensions and coordinates into a limited interval \([0.1; 0.9]\). Applying equation 6, we obtain a normalized vector as described next.

\[
x^n_a = 0.1 + 0.8 \left[ \frac{(x^a_m - x^a_{\min})}{(x^a_{\max} - x^a_{\min})} \right] \]

(6)

Where:
- \( x^n_a \) is the normalized value of variable \( a \) related to sample \( m \);
- \( x^a_m \) is the original value of variable \( a \) related to sample \( m \);
- \( x^a_{\max} \) is the maximum value for variable \( a \); and
- \( x^a_{\min} \) is the minimum value for variable \( a \).

Next, we processed the data related to vector \( \vec{AV} \) by using equation 3 and then equation 4 to obtain \( \vec{Y} \) vector.

In the sequence, training (75%) and test (25%) independent data sets were generated by a random and proportional selection. Next, three-layered structures with sigmoid activation functions in the neuron outputs were defined. Using these sets, we conducted four types of simulations, as previously defined on Table 1, by applying a backpropagation algorithm and considering a learning rate of 0.1. The networks were trained and Table 2 presents the best results reached for each type of simulation and attributes.
Table 2 – Recognition rates (%) by attributes and types of simulation

<table>
<thead>
<tr>
<th>Attributes</th>
<th>comfort</th>
<th>Frequency</th>
<th>Price</th>
<th>safety</th>
<th>Speed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type 1</td>
<td>33.33</td>
<td>55.00</td>
<td>0.00</td>
<td>80.00</td>
<td>38.46</td>
<td>53.73</td>
</tr>
<tr>
<td>Type 2</td>
<td>0.00</td>
<td>66.67</td>
<td>0.00</td>
<td>96.00</td>
<td>0.00</td>
<td>53.73</td>
</tr>
<tr>
<td>Type 3</td>
<td>0.00</td>
<td>61.11</td>
<td>0.00</td>
<td>80.00</td>
<td>30.77</td>
<td>52.24</td>
</tr>
<tr>
<td>Type 4</td>
<td>16.67</td>
<td>55.56</td>
<td>0.00</td>
<td>76.00</td>
<td>30.77</td>
<td>50.75</td>
</tr>
</tbody>
</table>

4.3 Analysis of the results

We notice that simulation Type 1 provided the best results among all (53.73%). This conclusion could not be reached only analyzing the total recognition rate since it is the same of simulation Type 2. So, we have to evaluate the results for each attribute and then it is clear that simulation Type 1 has a better performance than the others do. Despite the fact that the best recognition (96%) was reached for simulation Type 2 and attribute safety, NN’s capability of generalization was only reached for Type 1. This means that NN-Type 1 has the capability to better forecast all attributes and it can be used for all the rest of Taguatinga City.

The combination of individual and spatial characteristics can be pointed out as the main reason for this result. Simulation Type 1 combines socio-economic, demographic, user’s habits and spatial characteristics, which are normalized and coded. Especially spatial characteristics provide additional information for NN training in the sense that user’s characteristics are similar all over the case study area. Then, conditions of the urban environment are decisive to describe user’s preferences related to trip’s attributes.

Additionally to role of spatial characteristic, we also verify a small but important contribution reached from codification and normalization procedures. The application of these two pre-processing techniques has generated a considerable improvement in the NN. This can be explained by the fact that some variables are more suitable to codification than to normalization, as used in simulation Type 3 and 4.

Another point of interest on Table 2 is the high variation of recognition rates among the attributes. It is easily noticed that comfort, price and speed attributes are poorly represented, since the former reached just 33% (Type 1) of recognition while price was completely unpredicted. Meanwhile, safety and frequency attributes presented very high rates, indicating that the training data set was able to correctly express their nature.

There are some aspects that can help on the understanding of these results. Firstly, the composition of the training data set is examined. From 209 vectors in the training data set, comfort, frequency, price, safety and speed attributes are expressed by 18 (9%), 56 (27%), 17 (8%), 76 (36%) and 42 (20%) vectors, respectively. This shows a large concentration of vectors for safety attribute, while price attribute has very few samples, characterizing an imbalance training set problem which is well known when NN mainly models the dominant class.

In addition to the analysis of data set composition, we also have to consider the nature and origin of the collected data by Martins [5] as a crucial point.
Socio-economic, demographic and trip’s habit data were surveyed intending to apply traditional statistical analysis without regarding the spatial nature of the problem. Consequently, we observe that there was a limited concern about sample’s distribution on the space. For instance, the 276 samples can be grouped into 51 dwellings, i.e., the most part of the interviews were conducted in the same place. There were cases where 10 interviews were conducted in the same house. Obviously and unfortunately, this underestimates the real condition and needs of PBT’s users. Nevertheless, the NN was able to correctly forecast some of the attributes, expressing its potential for such a problem.

5 Conclusion

We initially presented the current situation of PBT in Brazil. In this situation, planning agencies require new instruments to develop new strategies for urban transportation services. These strategies need to be focused towards potential users, in opposition to previous notion that considered all population as homogeneous group without special requirements.

In an initial, but expected to be an efficient effort, a new conception and model to forecast user’s preferences were described. They intend to be innovative by associating technological tools (GIS and NN) to a modern conception (Marketing) to obtain fundamental information for PBT’s planning. It is expected that such modelling will contribute to reach user’s preferences through the simultaneously incorporation of individual and spatial characteristics inside an urban environment.

In this paper, we concentrated on the mathematical description of the proposed model. We defined a general framework intending to allow future improvements according to specific needs and situations. Therefore, our modelling description has much more to be explored from now on.

The case study showed, above all, the potential of the proposed model for a real application. Through out the simulations, it was noticed the influence of the incorporation of spatial characteristics in order to provide a better “learning” and “generalization” using NN. We also verified that pre-processing procedures (codification and normalization) have an impact on the results. Finally, we observe a strong indicator that the modelling performance would be better if the collection of data had considered more properly the spatial reality. It is well known that traditional statistical models have induced to the conception of restricted representation of urban space and dynamic due to limited attention to data collection activities.

Four main perspectives for improvements of this model can be highlighted. First of all, it is necessary to develop a specific survey devoted to a neural-spatial modelling in order to avoid the concentration of data in some attributes. The second perspective is related to the creation of methodology to incorporate the influence of the urban spatial structure and its diversity into the decision making process of PBT’s potential users. Next, it is fundamental to conceive a temporal-series analysis to evaluate changes along the time and their relationship with urban dynamic. Finally, new structures of NN must be simulated in order to
reach a dedicated processing, which has to be specifically devoted to forecast user's preferences.

Acknowledgments

The authors wish to thank the Brazilian Scientific and Technologic Development Agency (CNPq) and the Japanese Ministry of Education (Monbushou) for the scholarships that supported the development of this research.

References

